CSI Revisited: The Science of Forensic DNA Analysis

Michael L. Raymer
Wright State University - Main Campus, michael.raymer@wright.edu

Follow this and additional works at: https://corescholar.libraries.wright.edu/knoesis

Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation
https://corescholar.libraries.wright.edu/knoesis/927

This Presentation is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
CSI Revisited
The Science of Forensic DNA Analysis

Michael L. Raymer, Ph.D.
Growth in the importance of DNA

- Roughly 900,000 felony convictions per year in the U.S.
- DNA profiles are generated primarily for sexual offenses, murder, and assault
  - Often the key source of physical evidence
- The F.B.I. has established the CODIS database, with over 2 million DNA profiles
  - Allows “cold hit” searches for unresolved cases
DNA evidence misconceptions

- Everyone’s DNA profile is unique
- DNA testing is always an objective and scientific process
- DNA testing is infallible
- DNA evidence is carefully evaluated by both the prosecution and the defense

We’ve got him cold.
The science of DNA testing is sound

but

not all DNA testing is done scientifically
Background: DNA

- DNA is found in each human cell

<table>
<thead>
<tr>
<th>Type of Sample</th>
<th>Amount of DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood</td>
<td>30,000 ng/mL</td>
</tr>
<tr>
<td>1 cm² stain</td>
<td>200 ng</td>
</tr>
<tr>
<td>1 mm² stain</td>
<td>2 ng</td>
</tr>
<tr>
<td>Semen</td>
<td>250,000 ng/mL</td>
</tr>
<tr>
<td>Postcoital vaginal swab</td>
<td>0 – 3,000 ng</td>
</tr>
<tr>
<td>Hair</td>
<td></td>
</tr>
<tr>
<td>Plucked</td>
<td>1 – 750 ng/hair</td>
</tr>
<tr>
<td>Shed</td>
<td>1 – 12 ng/hair</td>
</tr>
<tr>
<td>Saliva</td>
<td>5,000 ng/mL</td>
</tr>
<tr>
<td>Urine</td>
<td>1 – 20 ng/mL</td>
</tr>
</tbody>
</table>
Background: DNA structure

DNA is a polymer of nucleotides
- Four building blocks: A, C, G, T
Background: DNA information content

- Most DNA (as much as 90%) is non-coding, or “junk” DNA
- More than 99% of the DNA is identical between any two humans
  - Regions of difference: “polymorphic”
- Changes to DNA are random, and usually bad

**Non-coding DNA exhibits higher polymorphism**
STRs

- **Short Tandem Repeat (STR)**
- Describes a type of DNA polymorphism in which:
  - a DNA sequence repeats
  - over and over again
  - and has a short (usually 4 base pair) repeat unit
- A length polymorphism – alleles differ in their length
  - 3 repeats: AATG AATG AATG
  - 4 repeats: AATG AATG AATG AATG
  - 5 repeats: AATG AATG AATG AATG AATG
  - 6 repeats: AATG AATG AATG AATG AATG AATG
13 CODIS core STR loci

- TPOX
- D3S1358
- D5S818
- FGA
- CSF1PO
- D8S1179
- D7S820
- TH01
- VWA
- D13S317
- D16S539
- D18S51
- D21S11
- AMEL
- AMEL
Short Tandem Repeats (STRs)

- The repeat region is variable between samples while the flanking regions where PCR primers bind are constant.

- AATG

- 7 repeats

- 8 repeats

- Homozygote = both alleles are the same length

- Heterozygote = alleles differ and can be resolved from one another
Extract and Purify DNA

- Add primers and other reagents
PCR Amplification

- DNA regions flanked by primers are amplified

Groups of amplified STR products are labeled with different colored dyes (blue, green, yellow)
Profiler Plus: After Amplification
The ABI 310 Genetic Analyzer:
Capillary Electrophoresis

- Amplified STR DNA injected onto column
- Electric current applied
- DNA pulled towards the positive electrode
- DNA separated out by size:
  - Large STRs travel slower
  - Small STRs travel faster
- Color of STR detected and recorded as it passes the detector
‘Nested’ STR alleles: Profiler Plus

Small

D3S1358
D8S1179
D5S818

Medium

vWA
D21S11
D13S317

Large

FGA
D18S51
D7S820

Blue

16,17
15,15
16,16

Green

16,16
28,29
8
12

Yellow

10,10
11,13
19
14

© M. Raymer, FBS
Profiler Plus: Raw data
• GENESCAN divides the raw data into a separate electropherogram for each color:
  • Blue
  • Green
  • Yellow
  • Red

- D3: 16, 17
- vWA: 15, 15
- FGA: 21, 23
- Amelogenin: X, Y
- D8: 16, 16
- D21: 28, 29
- D18: 14, 19
- D5: 8, 12
- D13: 11, 13
- D7: 10, 10
Reading an electropherogram

**Amelogenin**

**XX** = female

**XY** = male

**Peaks correspond to alleles**
### Allele Frequencies

**Locus** D3S1358  
**Race** Caucasian  
**N** = 203  

<table>
<thead>
<tr>
<th>Allele</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>0.12</td>
</tr>
<tr>
<td>13</td>
<td>0.12</td>
</tr>
<tr>
<td>14</td>
<td>0.140</td>
</tr>
<tr>
<td>15</td>
<td>0.026</td>
</tr>
<tr>
<td>16</td>
<td>0.003</td>
</tr>
<tr>
<td>17</td>
<td>0.040</td>
</tr>
<tr>
<td>18</td>
<td>0.103</td>
</tr>
<tr>
<td>19</td>
<td>0.012</td>
</tr>
</tbody>
</table>

**Locus** vWA  
**Race** Caucasian  
**N** = 196  

<table>
<thead>
<tr>
<th>Allele</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0.12</td>
</tr>
<tr>
<td>12</td>
<td>0.12</td>
</tr>
<tr>
<td>13</td>
<td>0.12</td>
</tr>
<tr>
<td>14</td>
<td>0.102</td>
</tr>
<tr>
<td>15</td>
<td>0.082</td>
</tr>
</tbody>
</table>

Statistical estimates: product rule

\[
0.222 \times 0.222 \times 2 = 0.1
\]
The product rule

Allele Frequencies

Locus D3S1358
Race Caucasian
(N = 203)

<table>
<thead>
<tr>
<th>Allele</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>0.012</td>
</tr>
<tr>
<td>13</td>
<td>0.012</td>
</tr>
<tr>
<td>14</td>
<td>0.140</td>
</tr>
<tr>
<td>15</td>
<td>0.246</td>
</tr>
<tr>
<td>16</td>
<td>0.222</td>
</tr>
<tr>
<td>17</td>
<td>0.222</td>
</tr>
<tr>
<td>18</td>
<td>0.163</td>
</tr>
<tr>
<td>19</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Locus vWA
Race Caucasian
(N = 196)

<table>
<thead>
<tr>
<th>Allele</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0.012</td>
</tr>
<tr>
<td>12</td>
<td>0.012</td>
</tr>
<tr>
<td>13</td>
<td>0.012</td>
</tr>
<tr>
<td>14</td>
<td>0.102</td>
</tr>
<tr>
<td>15</td>
<td>0.082</td>
</tr>
</tbody>
</table>

1 in 10 \times 1 in 111 \times 1 in 20 = 0.1

1 in 100 \times 1 in 14 \times 1 in 81

1 in 116 \times 1 in 17 \times 1 in 16

1 in 79,531,528,960,000,000,000

1 in 80 quadrillion
Cofiler

D3S1358

AMEL

TH01

TPOX

CSF1PO

D16S539

D7S820
Components of a DNA report

- The samples tested
  - Evidence samples (crime scene)
  - Reference samples (defendant, suspect)
- The lab doing the testing
- The test used:
  - Profiler Plus, Cofiler, Identifiler, mtDNA
- The analyst who did the testing
- Results and conclusions:
  - Table of alleles
  - Narrative conclusions
### Table of alleles

#### TABLE OF RESULTS

<table>
<thead>
<tr>
<th>ITEM</th>
<th>DESCRIPTION</th>
<th>D3S1358</th>
<th>vWA</th>
<th>FGA</th>
<th>AMEL</th>
<th>D8S1179</th>
<th>D21S11</th>
<th>D18S51</th>
<th>D5S818</th>
<th>D13S317</th>
<th>D7S820</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Reference From</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Victim</td>
<td>15,18</td>
<td>18,20</td>
<td>26,28</td>
<td>X,X</td>
<td>10,13</td>
<td>30,31</td>
<td>12,17</td>
<td>12,12</td>
<td>11,12</td>
<td>10,12</td>
</tr>
<tr>
<td>2</td>
<td>Reference From</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Defendant</td>
<td>15,16</td>
<td>15,16</td>
<td>19,26</td>
<td>X,Y</td>
<td>12,13</td>
<td>31,31</td>
<td>16,21</td>
<td>11,12</td>
<td>11,12</td>
<td>10&gt;11</td>
</tr>
<tr>
<td>3</td>
<td>Neck Swab</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(18)</td>
<td>15,16</td>
<td>15,16</td>
<td>19,26</td>
<td>X,Y</td>
<td>12,13</td>
<td>31,31</td>
<td>16,21</td>
<td>11,12</td>
<td>11,12</td>
<td>10,11</td>
</tr>
<tr>
<td></td>
<td>(28)</td>
<td>(18,20)</td>
<td>(28)</td>
<td></td>
<td></td>
<td>(10)</td>
<td>(30)</td>
<td>(12,17)</td>
<td>11,12</td>
<td>11,12</td>
<td>(12)</td>
</tr>
<tr>
<td>4</td>
<td>Chest Swab</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(17&lt;18)</td>
<td>15,16</td>
<td>15,16</td>
<td>19,26</td>
<td>X&gt;Y</td>
<td>12,13</td>
<td>31,31</td>
<td>16,21</td>
<td>11,12</td>
<td>11,12</td>
<td>10,11</td>
</tr>
<tr>
<td></td>
<td>(28)</td>
<td>(18,20)</td>
<td>(28)</td>
<td></td>
<td></td>
<td>(10)</td>
<td>(30)</td>
<td>(12,17)</td>
<td>11,12</td>
<td>11,12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Extraction Blanks</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Key:  
- **NA** = No activity.  
- () = Weak results for types in parenthesis.  
- > = Greater than.  
- < = Less than.  
- X,X = Female DNA.  
- X,Y = Male DNA.

- Some labs include more information than others
- Usually includes information about mixed samples
- May also include:  
  - Indication of low level results  
  - Indication of results not reported  
  - Relative amounts of different alleles (in mixed samples)
- No standard format
Narrative conclusions

CONCLUSIONS

1. The neck and chest swabs (items 3 and 4) have an elevated level of amylase 1 present in the extracts. These results strongly indicate saliva on the swabs.

2. The genetic marker results in the DNA extracted from the neck and chest swabs (items 3 and 4) are a mixture of at least two persons. The results indicate a major (or stronger donor) and a secondary (or weaker donor). [Defendant] is, in my opinion, the major DNA donor on items 3 and 4. Due to the presence of weak typing results at some loci, it is possible that minor components of the mixture have dropped out in the larger loci. As a result, [Victim] cannot be excluded as a contributor to the secondary DNA profile obtained from the neck and chest swabs (items 3 and 4). In addition, a weak amount of D3S1358 type 17 was detected on item 4 which could not have originated from [Victim] or [Defendant]. It is unclear as to whether this allele is artifactual in origin or from another donor.

- Indicates which samples match
- Includes a statistical estimate
- Identifies samples as mixed
- May include an ‘identity statement’ i.e., samples are from the same source to a scientific degree of certainty (FBI)
- May allude to problems (e.g. interpretative ambiguity, contamination)
Sources of ambiguity in STR interpretation

- Degradation
- Allelic dropout
- False peaks
- Mixtures
- Accounting for relatives
- Threshold issues -- marginal samples
When biological samples are exposed to adverse environmental conditions, they can become degraded

- Warm, moist, sunlight, time

Degradation breaks the DNA at random

Larger amplified regions are affected first

Classic ‘ski-slope’ electropherogram

Peaks on the right lower than peaks on the left
Allelic Dropout

- Peaks in evidence samples all very low
  - Mostly below 150 rfu
- Peaks in reference sample much higher
  - All well above 800 rfu
- At D13S817:
  - Reference sample: 8, 14
  - Evidence sample: 8, 8
- 14 allele has dropped out -- or has it?
- Tend to see with ‘marginal samples’
False peaks & machine problems

- **False peaks:**
  - Contamination
  - Dye blob
  - Electrical spikes
  - Pull-up

- **Machine problems:**
  - Noise
  - Baseline instability
  - Injection failures
## Summary Sheet

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>GT Results</th>
<th>D3</th>
<th>vWA</th>
<th>FGA</th>
<th>D16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane Doe Victim PP</td>
<td>GT Graph</td>
<td>14 (1079)</td>
<td>15 (926)</td>
<td>16 (926)</td>
<td>19 (664)</td>
</tr>
<tr>
<td>Jane Doe-C</td>
<td>GT Graph</td>
<td>14 (858)</td>
<td>15 (794)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>John Doe Defendant PP</td>
<td>GT Graph</td>
<td>14 (159)</td>
<td>14b (2820)</td>
<td>19 (58)</td>
<td></td>
</tr>
<tr>
<td>John Doe-C</td>
<td>GT Graph</td>
<td>14 (2168)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dress - Blood Stain Evidence 1 PP</td>
<td>GT Graph</td>
<td>14 (851)</td>
<td>15 (759)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dress - Blood Stain-C Evidence 1 CO</td>
<td>GT Graph</td>
<td>14 (422)</td>
<td>15 (380)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knife - Blood Stain Evidence 2 PP</td>
<td>GT Graph</td>
<td>14 (1161)</td>
<td>15 (1084)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knife - Blood Stain-C Evidence 2 CO</td>
<td>GT Graph</td>
<td>14 (1003)</td>
<td>15 (895)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Towel Evidence 3 PP</td>
<td>GT Graph</td>
<td>14 (1032)</td>
<td>15 (1089)</td>
<td>16 (822)</td>
<td>19 (690)</td>
</tr>
<tr>
<td>Towel-C Evidence 3 CO</td>
<td>GT Graph</td>
<td>14 (669)</td>
<td>15 (640)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Control Run 1 Positive PP run1</td>
<td>GT Graph</td>
<td>14 (1399)</td>
<td>15 (1282)</td>
<td>16 (104)</td>
<td>17 (1358)</td>
</tr>
<tr>
<td>Positive Control Run 1-C Positive CO run1</td>
<td>GT Graph</td>
<td>14 (1032)</td>
<td>15 (858)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Control Run 2 pos_cont_co run2</td>
<td>GT Graph</td>
<td>14 (989)</td>
<td>15 (871)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Control Run 3 pos_cont_pp run3</td>
<td>GT Graph</td>
<td>14 (1262)</td>
<td>15 (1129)</td>
<td>16 (94)</td>
<td>17 (1253)</td>
</tr>
<tr>
<td>Reagent Blank Run 1 BLANK run1</td>
<td>GT Graph</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reagent Blank Run 2 blank run2</td>
<td>GT Graph</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The * indicates that this peak may be involved in pullup...
A locus by locus description of issues that may warrant further review by an expert, including:

- Peak height imbalance
- Presence of a mixture
- Possible degradation
- Possible pullup
- Inconsistent results from multiple runs
- Problems with control runs and reagent blanks

The reference samples of the victim, "Jane Doe", and "Jane Doe-C", Jane Doe-C displays peak height imbalance at the locus CSF. The difference in the peak heights of the 13 and 11 alleles for the CSF locus (51 and 889, respectively) could be the result of a technical artifact (such as primer binding site mutations), or be evidence of more than one contributor to that sample.

Jane Doe is consistent with its source being a mixture of two or more individuals. Two loci, D3 (Allele 14 - 1079 RFUs, Allele 15 - 926 RFUs, Allele 16*a - 102 RFUs) and D21 (Allele 27 - 806 RFUs, Allele 32.2 - 695 RFUs, Allele 34.2 - 56 RFUs) appear to have more than two alleles. The additional peaks in this reference sample were found to be below the threshold of 150 RFUs, indicating that they are possibly caused by stochastic effects. Some additional peaks may be due to an uncommon technical artifact known as +4 stutter. A mixture in a reference sample could indicate that contamination has occurred.
What can be done to make DNA testing more objective?

- Distinguish between signal and noise
  - Deducing the number of contributors to mixtures
  - Accounting for relatives
Where do peak height thresholds come from (originally)?

- Applied Biosystems validation study of 1998
Where do peak height thresholds come from (originally)?

PCR products were examined on both the 377 DNA Sequencer and the 310 Genetic Analyzer. The results of 0.25 to 1.0 ng were clearly typable with peak heights of approximately 150 RFU and greater (data not shown). At 0.125 ng and less, the peak heights in both samples were not significantly above the background (< 150 RFU) or were undetectable. At 0.0313 ng specifically, peaks were extremely low or undetectable, and thus, DNA quantities as low as approximately 35 pg did not produce a typable result. Based on these results, we employed a peak height threshold of 150 RFU, below which peaks were interpreted with caution. Laboratories should determine a minimum peak height threshold for their instruments using high quality, single source genomic DNA samples which provides them with the desired sensitivity while not allowing for detection of low copy DNA. This is particularly important as the overall sensitivity of the assay may vary between laboratories.
Where do peak height thresholds come from?

- “Conservative” thresholds established during validation studies
- Eliminate noise (even at the cost of eliminating signal)
- Can arbitrarily remove legitimate signal
- Contributions to noise vary over time (e.g. polymer and capillary age/condition)
  - Analytical chemists use LOD and LOQ
Signal Measure

Saturation

Measured signal (In Volts/RFUS/etc)

\[ \mu_b + 10\sigma_b \]

Quantification limit

\[ \mu_b + 3\sigma_b \]

Detection limit

\[ \mu_b \]

Mean background

Signal

© M. Raymer, FBS
Opportunities to measure baseline
Control samples

- **Negative controls**: 5,932 data collection points (DCPs) per run ($\sigma = 131$ DCPs)
- **Reagent blanks**: 5,946 DCPs per run ($\sigma = 87$ DCPs)
- **Positive controls**: 2,415 DCP per run ($\sigma = 198$ DCPs)
- **DCP regions corresponding to size standards and 9947A peaks (plus and minus 55 DCPs to account for stutter in positive controls)** were masked in all colors
RFU levels at all non-masked data collection points
### Variation in baseline noise levels

<table>
<thead>
<tr>
<th></th>
<th>$\mu_b$</th>
<th>$\sigma_b$</th>
<th>$\mu_b + 3\sigma_b$</th>
<th>$\mu_b + 10\sigma_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive Control</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>6.7</td>
<td>6.9</td>
<td>27.4</td>
<td>75.7</td>
</tr>
<tr>
<td>Average</td>
<td>5.0</td>
<td>3.7</td>
<td>16.1</td>
<td>42.0</td>
</tr>
<tr>
<td>Minimum</td>
<td>3.7</td>
<td>2.4</td>
<td>10.9</td>
<td>27.7</td>
</tr>
<tr>
<td><strong>Negative Control</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>13.4</td>
<td>13.2</td>
<td>53.0</td>
<td>145.4</td>
</tr>
<tr>
<td>Average</td>
<td>5.4</td>
<td>3.9</td>
<td>17.1</td>
<td>44.4</td>
</tr>
<tr>
<td>Minimum</td>
<td>4.0</td>
<td>2.6</td>
<td>11.8</td>
<td>30.0</td>
</tr>
<tr>
<td><strong>Reagent Blank</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>6.5</td>
<td>11.0</td>
<td>39.5</td>
<td>116.5</td>
</tr>
<tr>
<td>Average</td>
<td>5.3</td>
<td>4.0</td>
<td>17.3</td>
<td>45.3</td>
</tr>
<tr>
<td>Minimum</td>
<td>4.0</td>
<td>2.6</td>
<td>11.8</td>
<td>30.0</td>
</tr>
<tr>
<td><strong>All three controls averaged</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>7.1</td>
<td>7.3</td>
<td>29.0</td>
<td>80.1</td>
</tr>
<tr>
<td>Average</td>
<td>5.2</td>
<td>3.9</td>
<td>16.9</td>
<td>44.2</td>
</tr>
<tr>
<td>Minimum</td>
<td>3.9</td>
<td>2.5</td>
<td>11.4</td>
<td>28.9</td>
</tr>
</tbody>
</table>

Average ($\mu_b$) and standard deviation ($\sigma_b$) values with corresponding LODs and LOQs from positive, negative and reagent blank controls in 50 different runs. BatchExtract: ftp://ftp.ncbi.nlm.nih.gov/pub/forensics/
Two reference samples in a 1:10 ratio (male:female). Three different thresholds are shown: 150 RFU (red); LOQ at 77 RFU (blue); and LOD at 29 RFU (green).
Familial searching

- Database search yields a close but imperfect DNA match
- Can suggest a relative is the true perpetrator
- Great Britain performs them routinely
- Reluctance to perform them in US since 1992 NRC report
- Current CODIS software cannot perform effective searches
Three approaches to familial searches

- Search for rare alleles (inefficient)
- Count matching alleles (arbitrary)
- Likelihood ratios with kinship analyses
Pair-wise similarity distributions

- Number of pairwise shared alleles
- Percent of total (%)

Graph showing distributions for:
- Randomized Individuals
- Simulated Cousins
- Simulated Siblings
Is the true DNA match a relative or a random individual?

- Given a closely matching profile, who is more likely to match, a relative or a randomly chosen, unrelated individual?

- Use a likelihood ratio

\[ LR = \frac{P(E \mid \text{relative})}{P(E \mid \text{random})} \]
Is the true DNA match a relative or a random individual?

- What is the likelihood that a relative of a single initial suspect would match the evidence sample perfectly?

- What is the likelihood that a single randomly chosen, unrelated individual would match the evidence sample perfectly?

\[
LR = \frac{P(E \mid \text{relative})}{P(E \mid \text{random})}
\]
Probabilities of siblings matching at 0, 1 or 2 alleles

\[ P(E | sib) = \begin{cases} 
\frac{P_a \cdot P_b \cdot HF}{4}, & \text{if } shared = 0 \\
\frac{P_b + P_a \cdot P_b \cdot HF}{4}, & \text{if } shared = 1 \\
\frac{1 + P_a + P_b + P_a \cdot P_b \cdot HF}{4}, & \text{if } shared = 2 
\end{cases} \]

HF = 1 for homozygous loci and 2 for heterozygous loci; \( P_a \) is the frequency of the allele shared by the evidence sample and the individual in a database.
Probabilities of parent/child matching at 0, 1 or 2 alleles

\[
P(E \mid \text{parent/child}) = \begin{cases} 
0, & \text{if } \text{shared} = 0 \\
\frac{P_b}{2}, & \text{if } \text{shared} = 1 \\
\frac{P_a + P_b}{2}, & \text{if } \text{shared} = 2
\end{cases}
\]

HF = 1 for homozygous loci and 2 for heterozygous loci; \(P_a\) is the frequency of the allele shared by the evidence sample and the individual in a database.
Other familial relationships

Cousins:

\[
P(E \mid cousins) = \begin{cases} 
  \frac{6 \cdot P_a \cdot P_b \cdot HF}{8}, & \text{if } shared = 0 \\
  \frac{P_b + 6 \cdot P_a \cdot P_b \cdot HF}{8}, & \text{if } shared = 1 \\
  \frac{P_a + P_b + 6 \cdot P_a \cdot P_b \cdot HF}{8}, & \text{if } shared = 2 
\end{cases}
\]

\[
P(E \mid GG / AUNN / HS) = \begin{cases} 
  \frac{2 \cdot P_a \cdot P_b \cdot HF}{4}, & \text{if } shared = 0 \\
  \frac{P_b + 2 \cdot P_a \cdot P_b \cdot HF}{4}, & \text{if } shared = 1 \\
  \frac{P_a + P_b + 2 \cdot P_a \cdot P_b \cdot HF}{4}, & \text{if } shared = 2 
\end{cases}
\]

HF = 1 for homozygous loci and 2 for heterozygous loci; \( P_a \) is the frequency of the allele shared by the evidence sample and the individual in a database.
Familial search experiment

- Randomly pick related pair or unrelated pair from a synthetic database
- Choose one profile to be evidence and one profile to be initial suspect
- Test hypothesis:
  - $H_0$: A relative is the source of the evidence
  - $H_A$: An unrelated person is the source of the evidence

Hypothesis testing: LR threshold of 1 with prior odds of 1

<table>
<thead>
<tr>
<th>Decision</th>
<th>True state</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Evidence from Unrelated individual</td>
</tr>
<tr>
<td>Evidence from unrelated individual</td>
<td>~ 98% [Correct decision]</td>
</tr>
<tr>
<td>Evidence from sibling</td>
<td>~ 2% [Type I error; false positive]</td>
</tr>
</tbody>
</table>
Two types of errors

- False positives (Type I): an initial suspect’s family is investigated even though an unrelated individual is the actual source of the evidence sample.

- False negatives (Type II): an initial suspect’s family is not be investigated even though a relative really is the source of the evidence sample.

- A wide net (low LR threshold) catches more criminals but comes at the cost of more fruitless investigations.
Type I and II errors with prior odds of 1
Is the true DNA match a relative or a random individual?

- What is the likelihood that a close relative of a single initial suspect would match the evidence sample perfectly?

- What is the likelihood that a single randomly chosen, unrelated individual would match the evidence sample perfectly?

\[ LR = \frac{P(E \mid \text{relative})}{P(E \mid \text{random})} \]
Is the true DNA match a relative or a random individual?

- What is the likelihood that the source of the evidence sample was a relative of an initial suspect?

\[
P(sib \mid E) = \frac{P(E \mid sib) \cdot P(sib)}{P(E \mid sib) \cdot P(sib) + P(E \mid random) \cdot P(random)}
\]

\[
P(sib) = \frac{s}{\text{popsize}}
\]

\[
P(random) = \frac{\text{popsize} - s}{\text{popsize}}
\]
Is the true DNA match a relative or a random individual?

- This more difficult question is ultimately governed by two considerations:
  - What is the size of the alternative suspect pool?
  - What is an acceptable rate of false positives?

$$LR = \frac{P(E \mid sib)}{P(E \mid random)}$$
Pair-wise similarity distributions

Number of pairwise shared alleles

Percent of total (%)
How well does an LR approach perform relative to alternatives?

- Low-stringency CODIS search identifies all 10,000 parent-child pairs (but only 1,183 sibling pairs and less than 3% of all other relationships and a high false positive rate)

- Moderate and high-stringency CODIS searches failed to identify any pairs for any relationship

- An allele count-threshold (set at 20 out of 30 alleles) identifies 4,233 siblings and 1,882 parent-child pairs (but fewer than 70 of any other relationship and with no false positives)
How well does an LR approach perform relative to alternatives?

- LR set at 1 identifies >99% of both sibling and parent-child pairs (with false positive rates of 0.01% and 0.1%, respectively)
- LR set at 10,000 identifies 64% of siblings and 56% of parent-child pairs (with no false positives)
- Use of non-cognate allele frequencies results in an increase in false positives and a decrease in true positives (that are largely offset by either a ceiling or consensus approach)
Introduction to Mixtures

- Mixtures can exhibit up to two peaks per contributor at any given locus.
- Mixtures can exhibit as few as 1 peak at any given locus (regardless of the number of contributors).
Determining if two genotypes could be contributors is relatively easy.

Possible contributors to a mixture:

- **D3 locus genotype**
  - Individual #1: 15, 18
  - Individual #2: 14, 18
  - Mixture: 14, 15, 18

But beware – the opposite is not true.
Introduction to Mixtures

- Determining what genotypes created the mixture is non-trivial

D3 locus genotype
Mixture: 14, 15, 18

Option #1
Individual A: 15, 18
Individual B: 14, 18

Option #2
Individual B: 14, 18
Individual C: 15, 15

Option #3
Individual #D: 14, 15
Individual #E: 18, 18

Option #4
Individual #A: 15, 18
Individual #F: 14, 14
Even determining the number of contributors is non-trivial.

D3 locus genotype
Mixture: 14, 15, 18

Another Option
Individual C: 15, 15
Individual D: 14, 15
Individual E: 18, 18

There is no “hard” mathematical upper bound to the number of contributors possible.
Usually the victim’s genotype is known, but this does not always make the defendant’s genotype clear.

**D3 locus genotype**

- **Mixture:** 14, 15, 18
- **Victim:** 14, 15

**Possible genotypes for a single perpetrator:**

- Individual C: 14, 18
- Individual D: 15, 18
- Individual E: 18, 18
- Individual F: 14, 14 ?
The large number of potential genotypes consistent with the mixture allows for a VERY wide net to be cast

- This greatly increases the likelihood of accusing an innocent suspect, particularly in database trawls
- This is generally not reflected in the statistics reported by the DNA testing laboratory

- Case History: Sutton
Making sense of mixtures

- There are two major open research areas:
  - Determining the most likely number of contributors
  - Determining the genotypes of each contributor

- Factors that can aid in deconvolution
  - Mixture ratios
  - Peak height additivity

- Factors that can greatly complicate deconvolution results
  - Allowing alleles to be discarded as artifacts ("analyst’s discretion")
Different individuals may contribute different “amounts” of DNA to the mixture. This difference should be reflected (relatively uniformly) throughout the entire sample.
Peak height additivity

- Assume one individual contributes an amount of DNA that measured at n RFUs
- Assume a second individual contributes an amount of DNA that measures at m RFUs
- In a two person mixture, any allele which they share should measure at roughly n + m RFUs
Evidence of additivity

Relationship between the smaller and larger peaks in heterozygous loci of reference samples.
Making sense of mixtures

- There are two major research areas:
  - **Determining the most likely number of contributors**
  - Determining the genotypes of each contributor

- How can we determine the mostly likely number of contributors?
  - We (Paoletti et al.) create mixtures from an existing database in order to determine how often the actual number of contributors differs from the perceived number of contributors.
  - The Minnesota BCA database uses twelve (12) loci
<table>
<thead>
<tr>
<th>BCA ID#</th>
<th>D3S1358</th>
<th>vWA</th>
<th>FGA</th>
<th>THO1</th>
<th>TPOX</th>
<th>CSF1PO</th>
<th>D5S818</th>
<th>D13S317</th>
<th>D7S820</th>
<th>D8S1179</th>
<th>D21S1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PB0005</td>
<td>17,18</td>
<td>16,16</td>
<td>21,24</td>
<td>6,8</td>
<td>10,11</td>
<td>11,12</td>
<td>12,14</td>
<td>11,12</td>
<td>8,10</td>
<td>13,14</td>
<td>29,29</td>
</tr>
<tr>
<td>PH0070</td>
<td>15,17</td>
<td>16,17</td>
<td>21,25</td>
<td>7,7</td>
<td>10,11</td>
<td>11,12</td>
<td>11,12</td>
<td>11,12</td>
<td>8,10</td>
<td>13,14</td>
<td>29,2,</td>
</tr>
<tr>
<td>PH0138</td>
<td>17,17</td>
<td>14,16</td>
<td>24,25</td>
<td>7,8</td>
<td>11,11</td>
<td>10,11</td>
<td>11,11</td>
<td>10,10</td>
<td>10,11</td>
<td>14,14</td>
<td>29,30</td>
</tr>
<tr>
<td>Mixture1</td>
<td>15,17,18</td>
<td>14,16,17</td>
<td>21,24,25</td>
<td>6,7,8</td>
<td>10,11</td>
<td>10,11,12</td>
<td>11,12,14</td>
<td>10,11,12</td>
<td>8,10,11</td>
<td>13,14</td>
<td>29,29</td>
</tr>
<tr>
<td>PB0155</td>
<td>16,17</td>
<td>16,16</td>
<td>24,24</td>
<td>8,9.3</td>
<td>10,11</td>
<td>11,12</td>
<td>11,13</td>
<td>12,12</td>
<td>10,11</td>
<td>12,15</td>
<td>29,29</td>
</tr>
<tr>
<td>PH0014</td>
<td>17,17</td>
<td>17,18</td>
<td>19,22</td>
<td>6,9.3</td>
<td>11,12</td>
<td>12,12</td>
<td>9,9</td>
<td>11,11</td>
<td>13,15</td>
<td>28,29</td>
<td></td>
</tr>
<tr>
<td>PN0166</td>
<td>15,16</td>
<td>17,17</td>
<td>19,22</td>
<td>9.3,9.3</td>
<td>11,11</td>
<td>12,13</td>
<td>11,11</td>
<td>12,13</td>
<td>9,10</td>
<td>12,12</td>
<td>30,30</td>
</tr>
<tr>
<td>Mixture2</td>
<td>15,16,17</td>
<td>16,17,18</td>
<td>19,22,24</td>
<td>6,8,9.3</td>
<td>10,11,12</td>
<td>11,12,13</td>
<td>11,13</td>
<td>9,12,13</td>
<td>9,10,11</td>
<td>12,13,15</td>
<td>28,29</td>
</tr>
<tr>
<td>PB0022</td>
<td>15,16</td>
<td>15,16</td>
<td>22,23</td>
<td>7,7</td>
<td>9,11</td>
<td>11,12</td>
<td>11,12</td>
<td>14,14</td>
<td>11,12</td>
<td>14,15</td>
<td>32,2,</td>
</tr>
<tr>
<td>PH0078</td>
<td>15,17</td>
<td>15,15</td>
<td>23,24</td>
<td>7,8</td>
<td>10,10</td>
<td>11,12</td>
<td>11,12</td>
<td>13,13</td>
<td>10,10</td>
<td>13,13</td>
<td>28,28</td>
</tr>
<tr>
<td>PH0146</td>
<td>17,17</td>
<td>16,16</td>
<td>24,24</td>
<td>8,9.3</td>
<td>9,11</td>
<td>10,12</td>
<td>12,12</td>
<td>8,8</td>
<td>10,12</td>
<td>13,14</td>
<td>28,32</td>
</tr>
<tr>
<td>Mixture3</td>
<td>15,16,17</td>
<td>15,16</td>
<td>22,23,24</td>
<td>7,8,9.3</td>
<td>9,10,11</td>
<td>10,11,12</td>
<td>11,12</td>
<td>8,13,14</td>
<td>10,11,12</td>
<td>13,14,15</td>
<td>28,32</td>
</tr>
<tr>
<td>PB0024</td>
<td>17,18</td>
<td>16,18</td>
<td>22,24</td>
<td>7,8</td>
<td>6,9</td>
<td>10,11</td>
<td>11,11</td>
<td>9,12</td>
<td>8,10</td>
<td>15,15</td>
<td>29,29</td>
</tr>
<tr>
<td>PH0067</td>
<td>17,18</td>
<td>16,19</td>
<td>22,24</td>
<td>7,8</td>
<td>11,11</td>
<td>10,10</td>
<td>12,13</td>
<td>11,12</td>
<td>8,8</td>
<td>12,13</td>
<td>29,30</td>
</tr>
<tr>
<td>PB0111</td>
<td>15,18</td>
<td>16,16</td>
<td>23,24</td>
<td>8,9.3</td>
<td>6,9</td>
<td>10,11</td>
<td>11,12</td>
<td>11,12</td>
<td>10,12</td>
<td>12,15</td>
<td>30,31</td>
</tr>
<tr>
<td>Mixture4</td>
<td>15,17,18</td>
<td>16,18,19</td>
<td>22,23,24</td>
<td>7,8,9.3</td>
<td>6,9,11</td>
<td>10,11</td>
<td>11,12,13</td>
<td>9,11,12</td>
<td>8,10,12</td>
<td>12,13,15</td>
<td>29,30</td>
</tr>
<tr>
<td>PB0024</td>
<td>17,18</td>
<td>16,18</td>
<td>22,24</td>
<td>7,8</td>
<td>6,9</td>
<td>10,11</td>
<td>11,11</td>
<td>9,12</td>
<td>8,10</td>
<td>15,15</td>
<td>29,29</td>
</tr>
<tr>
<td>PH0075</td>
<td>16,18</td>
<td>16,16</td>
<td>22,24</td>
<td>9.3,9.3</td>
<td>8,8</td>
<td>7,10</td>
<td>8,11</td>
<td>11,11</td>
<td>8,8</td>
<td>14,14</td>
<td>29,32</td>
</tr>
<tr>
<td>PC0090</td>
<td>16,17</td>
<td>14,18</td>
<td>22,25</td>
<td>7,8</td>
<td>8,8</td>
<td>10,11</td>
<td>12,12</td>
<td>11,11</td>
<td>8,12</td>
<td>12,15</td>
<td>29,30</td>
</tr>
<tr>
<td>Mixture5</td>
<td>16,17,18</td>
<td>14,16,18</td>
<td>22,24,25</td>
<td>7,8,9.3</td>
<td>6,8,9</td>
<td>7,10,11</td>
<td>8,11,12</td>
<td>9,11,12</td>
<td>8,10,12</td>
<td>12,14,15</td>
<td>29,30</td>
</tr>
<tr>
<td>PB0030</td>
<td>14,16</td>
<td>15,15</td>
<td>22,22</td>
<td>7,7</td>
<td>8,9</td>
<td>11,11</td>
<td>11,13</td>
<td>12,13</td>
<td>10,11</td>
<td>14,16</td>
<td>28,29</td>
</tr>
<tr>
<td>PH0055</td>
<td>16,16</td>
<td>16,18</td>
<td>24,24</td>
<td>7,9</td>
<td>8,11</td>
<td>11,12</td>
<td>12,14</td>
<td>8,11</td>
<td>13,14</td>
<td>28,29</td>
<td></td>
</tr>
<tr>
<td>PN0108</td>
<td>15,16</td>
<td>18,18</td>
<td>22,23</td>
<td>9.3,9.3</td>
<td>11,11</td>
<td>11,11</td>
<td>12,14</td>
<td>8,8</td>
<td>14,16</td>
<td>29,30</td>
<td></td>
</tr>
<tr>
<td>Mixture6</td>
<td>14,15,16</td>
<td>15,16,18</td>
<td>22,23,24</td>
<td>7,9,9.3</td>
<td>8,9,11</td>
<td>11,12</td>
<td>11,12,13</td>
<td>12,13,14</td>
<td>8,10,11</td>
<td>13,14,16</td>
<td>28,29</td>
</tr>
</tbody>
</table>
All 3-way MN BCA mixtures

- There are 45,139,896 possible different 3-person mixtures of the 648 individuals in the MN BCA database

<table>
<thead>
<tr>
<th>Maximum # of alleles observed</th>
<th># of occurrences</th>
<th>As Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>310</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>2,498,139</td>
<td>5.53%</td>
</tr>
<tr>
<td>5</td>
<td>29,938,777</td>
<td>66.32%</td>
</tr>
<tr>
<td>6</td>
<td>12,702,670</td>
<td>28.14%</td>
</tr>
</tbody>
</table>

- 6% of three contributors mixtures “look like” two contributors
All 3-way MN BCA mixtures

- What if “analyst’s discretion” is invoked exactly once (at the “worst” locus)

<table>
<thead>
<tr>
<th>Maximum # of alleles observed</th>
<th># of occurrences</th>
<th>As Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>310</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>8,151</td>
<td>0.02%</td>
</tr>
<tr>
<td>4</td>
<td>2,498,139</td>
<td>5.53%</td>
</tr>
<tr>
<td></td>
<td>11,526,219</td>
<td>25.53%</td>
</tr>
<tr>
<td>5</td>
<td>29,938,777</td>
<td>66.32%</td>
</tr>
<tr>
<td></td>
<td>32,078,976</td>
<td>71.01%</td>
</tr>
<tr>
<td>6</td>
<td>12,702,670</td>
<td>28.14%</td>
</tr>
<tr>
<td></td>
<td>1,526,550</td>
<td>3.38%</td>
</tr>
</tbody>
</table>

- 26% of three contributors mixtures “look like” two contributors
### All 4-way MN BCA mixtures

<table>
<thead>
<tr>
<th>Maximum # of alleles observed</th>
<th># of occurrences</th>
<th>As Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>42,923</td>
<td>0.07%</td>
</tr>
<tr>
<td></td>
<td>731,947</td>
<td>1.25%</td>
</tr>
<tr>
<td>5</td>
<td>9,365,770</td>
<td>15.03%</td>
</tr>
<tr>
<td></td>
<td>30,471,965</td>
<td>52.18%</td>
</tr>
<tr>
<td>6</td>
<td>34,067,153</td>
<td>58.32%</td>
</tr>
<tr>
<td></td>
<td>25,872,024</td>
<td>44.29%</td>
</tr>
<tr>
<td>7</td>
<td>13,719,403</td>
<td>23.49%</td>
</tr>
<tr>
<td></td>
<td>1,328,883</td>
<td>2.28%</td>
</tr>
<tr>
<td>8</td>
<td>1,214,261</td>
<td>2.08%</td>
</tr>
<tr>
<td></td>
<td>4,695</td>
<td>0.01%</td>
</tr>
</tbody>
</table>

- 73% of four contributors mixtures “look like” three contributors
## All 4-way MN BCA mixtures

<table>
<thead>
<tr>
<th>Maximum # of alleles observed</th>
<th># of occurrences</th>
<th>As Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>42,923</td>
<td>0.07%</td>
</tr>
<tr>
<td></td>
<td>731,947</td>
<td>1.25%</td>
</tr>
<tr>
<td>5</td>
<td>9,365,770</td>
<td>15.03%</td>
</tr>
<tr>
<td></td>
<td>30,471,965</td>
<td>52.18%</td>
</tr>
<tr>
<td>6</td>
<td>34,067,153</td>
<td>58.32%</td>
</tr>
<tr>
<td></td>
<td>25,872,024</td>
<td>44.29%</td>
</tr>
<tr>
<td>7</td>
<td>13,719,403</td>
<td>23.49%</td>
</tr>
<tr>
<td></td>
<td>1,328,883</td>
<td>2.28%</td>
</tr>
<tr>
<td>8</td>
<td>1,214,261</td>
<td>2.08%</td>
</tr>
<tr>
<td></td>
<td>4,695</td>
<td>0.01%</td>
</tr>
</tbody>
</table>

- 96% of four contributors mixtures “look like” three contributors when one locus can be dropped from consideration.
Removing possible relationships

- Redistribute alleles at each locus randomly
- New database of “synthetic” unrelated individuals with the same allele frequencies

<table>
<thead>
<tr>
<th>Individual</th>
<th>vWA Original</th>
<th>vWA Redistributed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18,19</td>
<td>15,18</td>
</tr>
<tr>
<td>2</td>
<td>18,18</td>
<td>18,18</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>648</td>
<td>14,15</td>
<td>14,19</td>
</tr>
</tbody>
</table>
### 3-way mixtures with all 12 loci

<table>
<thead>
<tr>
<th>Maximum # of alleles observed in a 3-person mixture</th>
<th># of occurrences</th>
<th>Percent of mixtures</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>310</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>2,498,139</td>
<td>5.53%</td>
</tr>
<tr>
<td>5</td>
<td>29,938,777</td>
<td>66.32%</td>
</tr>
<tr>
<td>6</td>
<td>12,702,670</td>
<td>28.14%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Maximum # of alleles observed in a 3-person mixture</th>
<th># of occurrences</th>
<th>Percent of mixtures</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>139.4</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>2,233,740.8</td>
<td>4.95%</td>
</tr>
<tr>
<td>5</td>
<td>29,829,482.0</td>
<td>66.08%</td>
</tr>
<tr>
<td>6</td>
<td>13,076,533.8</td>
<td>28.97%</td>
</tr>
</tbody>
</table>
How many loci until 4-way mixture doesn’t look like a 3-way mixture?

- Redistribute alleles across all individuals (by locus) and add to database

© M. Raymer, FBS
What if contributors are related?

- Clearly, determining the number of contributors to a DNA mixture is difficult when the contributors are unrelated.

- How much harder does it become when they are related?
Virtual families

- Parents randomly chosen from unrelated (randomized) database
- Random mating
- Creates databases of grandparents, parents, and grandchildren
Distributions of shared alleles

Number of Shared Alleles vs Percent of Total

- **Unrelated**
- **Cousins**
- **Siblings**

© M. Raymer, FBS
Likelihoods of shared alleles

Number of Shared Alleles

Likelihood

Unrelated
Cousins
Siblings
Analysis of Allele Sharing

- Clearly, it is difficult to definitively assign the number of contributors to a mixture.
- This difficulty must be fairly reported in random probability match statistics in order for such statistics to remain objective.
- Analyst discretion should be invoked cautiously, and always carefully double-checked for error.
- Likelihoods allow a analyst to infer the possible relationship between two individuals.
Mixture Deconvolution

- Even when the number of contributors is known (or assumed), separating mixtures into their components can be difficult.
Most methods start by inferring the mixture ratio:

Simple example: All loci heterozygous, two contributors
Minimal Basic Assumptions

- A primary assumption of all methods is peak additivity.
- Most labs assume peaks from the same source will vary by $\leq 30\%$. 
Objectives

- Start with **simple assumptions:**
  - Additivity with constant variance: $c$
  - Peaks below a minimum threshold (often 50 or 150 RFU) are not observable
  - Peaks above the saturation threshold (often 4000 RFU) are not measurable

- Obtain *provably correct* deconvolution where possible

- Identify when this is not possible
Method

- Assume the number of contributors
- Enumerate all possible mixture contributor combinations
- Determine which pairs of profiles contain peaks in balance
Example: assume two contributors, four peaks:

- For this locus, and $c = 1.3$, the combination (P1, P3) is out of balance because:

$$180 \times c < 2030$$
Example: Mixture of four peaks

<table>
<thead>
<tr>
<th>Contributor 1</th>
<th>Contributor 2</th>
<th>Mixture Condition 1</th>
<th>Mixture Condition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>P4 P3</td>
<td>P2 P1</td>
<td>P4 ≤ cP3</td>
<td>P2 ≤ cP1</td>
</tr>
<tr>
<td>P4 P2</td>
<td>P3 P1</td>
<td>P4 ≤ cP2</td>
<td>P3 ≤ cP1</td>
</tr>
<tr>
<td>P4 P1</td>
<td>P3 P2</td>
<td>P4 ≤ cP1</td>
<td>P3 ≤ cP2</td>
</tr>
</tbody>
</table>

- \( P4 \geq P3 \geq P2 \geq P1 \geq \text{Min. Threshold} \)
○ If only one row is satisfied, then the genotypes can be **unambiguously** and **provably** determined
Example: In the sweet spot

<table>
<thead>
<tr>
<th>Contributor 1</th>
<th>Contributor 2</th>
<th>Mixture Condition 1</th>
<th>Mixture Condition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>P4 P3</td>
<td>P2 P1</td>
<td>P4 \leq cP3</td>
<td>P2 \leq cP1</td>
</tr>
<tr>
<td>P4 P2</td>
<td>P3 P1</td>
<td>P4 \leq cP2</td>
<td>P3 \leq cP1</td>
</tr>
<tr>
<td>P4 P1</td>
<td>P3 P2</td>
<td>P4 \leq cP1</td>
<td>P3 \leq cP2</td>
</tr>
</tbody>
</table>

- P4 > cP2
  so we can’t have (P4, P2)
- P4 > cP1
  so we can’t have (P4, P1)
Example: Ambiguous Locus

○ P2 is within c of both P1 and P4, so we can have
  • (P1,P3) (P2,P4), or
  • (P1,P2) (P3,P4)

○ P4 cannot pair with P1
Example: No row satisfied

- P4 (for example) cannot pair with any other peak
- One of our assumptions (c or the number of contributors) is incorrect

<table>
<thead>
<tr>
<th>Contributor 1</th>
<th>Contributor 2</th>
<th>Mixture Condition 1</th>
<th>Mixture Condition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>P4 P3</td>
<td>P2 P1</td>
<td>P4 ≤ cP3</td>
<td>P2 ≤ cP1</td>
</tr>
<tr>
<td>P4 P2</td>
<td>P3 P1</td>
<td>P4 ≤ cP2</td>
<td>P3 ≤ cP1</td>
</tr>
<tr>
<td>P4 P1</td>
<td>P3 P2</td>
<td>P4 ≤ cP1</td>
<td>P3 ≤ cP2</td>
</tr>
<tr>
<td>Contributor 1</td>
<td>Contributor 2</td>
<td>Mixture Condition 1</td>
<td>Mixture Condition 2</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------</td>
<td>-------------------------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>P3 P3</td>
<td>P2 P1</td>
<td>None (homozygote)</td>
<td>$P_2 \leq c \times P_1$</td>
</tr>
<tr>
<td>P3 P2</td>
<td>P3 P1</td>
<td>$P_3 \leq c \times (P_2+P_1)$</td>
<td>$P_3 \geq (1/c) \times (P_2+P_1)$</td>
</tr>
<tr>
<td>P3 P2</td>
<td>P2 P1</td>
<td>$P_2 \leq c \times (P_3+P_1)$</td>
<td>$P_2 \geq (1/c) \times (P_3+P_1)$</td>
</tr>
<tr>
<td>P3 P2</td>
<td>P1 Pmpht</td>
<td>$P_3 \leq c \times P_2$</td>
<td>$P_1 \leq c \times P_{mpt}$</td>
</tr>
<tr>
<td>P3 P2</td>
<td>P1 P1</td>
<td>$P_3 \leq c \times P_2$</td>
<td>None</td>
</tr>
<tr>
<td>P3 P1</td>
<td>P2 Pmpht</td>
<td>$P_3 \leq c \times P_1$</td>
<td>$P_2 \leq c \times P_{mpt}$</td>
</tr>
<tr>
<td>P3 P1</td>
<td>P2 P2</td>
<td>$P_3 \leq c \times P_1$</td>
<td>None</td>
</tr>
<tr>
<td>P3 P1</td>
<td>P2 P1</td>
<td>$P_1 \leq c \times (P_3+P_2)$</td>
<td>$P_1 \geq (1/c) \times (P_3+P_2)$</td>
</tr>
<tr>
<td>P3 Pmpht</td>
<td>P2 P1</td>
<td>$P_3 \leq c \times P_{mpt}$</td>
<td>$P_2 \leq c \times P_1$</td>
</tr>
</tbody>
</table>
Advantages of the method

- If you accept the simple assumptions, the resulting mixture interpretations directly follow
- Interprets mixtures on a locus by locus basis
- Does not interpret ambiguous loci
Future work

- Mixture ratio can be inferred only from unambiguous loci, and then applied to perform an more aggressive interpretation of the ambiguous loci when desired.
- Confidence values can be applied to the more aggressively interpreted positions.
Acknowledgements

- **Research Students**
  - David Paoletti (analysis of allele sharing)
  - Jason Gilder (data collection, additivity study, mixture deconvolution)

- **Faculty**
  - Travis Doom
  - Dan Krane
  - Michael Raymer